Automated Keyword Extraction of Learning Materials Using Semantic Relations

1. INTRODUCTION

The poster will present our on-going research, which will develop new algorithms to automatically generate keywords from online documents that describes lesson plans in mathematics and science. The motivations for improving the current keyword extraction mechanism are twofold:

- Feedback from our previous study (described below) showed that the keyword extraction was the least satisfying component of our automatic metadata extraction mechanisms to the users.
- Our data indicated that human annotators often assigned keywords to a document that do not appear in the document, which were impossible for the current keyword extraction mechanism to generate.

Building upon TextRank by Mihalcea and Tarau [4], our approach is to use a graph-based algorithm to rank keywords, based on semantic relationships.

2. BACKGROUND

A team comprised of the Center for Natural Language Processing (CNLP) at Syracuse University and the Digital Learning Sciences (DLS) at the University Corporation for Atmospheric Research recently completed a project that integrated many digital library tools into one, which is called Metadata Assignment and Search Tool (MAST). [1] This tool enables libraries and museums to efficiently describe and disseminate their digital materials by 1) automatically generating metadata to assist the cataloger; 2) assisting in assigning educational standards to learning materials; and 3) customizing their workflows and collection management. Previous versions of these tools are deployed in the National Science Digital Library (NSDL) project to assist catalogers in adding materials to the online digital collection.

[1]The project is funded by the Institute of Museum and Library Services (IMLS).

The automatic metadata assignment uses Natural Language Processing technologies to process the text of the online documents, in html or pdf formats, and produces metadata elements for the Dublin Core + GEM fields. These fields include general elements such as title, description, subject fields and contributors, as well as educational fields such as audience, instructional methods and grade level. The automatically generated fields are presented to the collection cataloger, who may correct or add to them. In user study tests with a group of managers, curators and directors representing both museums and libraries, their reviews of the process of cataloging with automatic metadata suggestion and managed workflows were generally favorable and enthusiastic. However, part of the feedback received from this group was that while almost all of the automatic metadata was helpful in cataloging, subject terms were not. Thus our current research lies in improving the automatic generation of these subject terms, which we will call keyword extraction, in keeping with the terminology used by other researchers in this area.

3. RELATED WORK

The idea of using the graph-based approach for information retrieval systems appeared in the early days of the research, such as THOMAS, a human-machine dialogue system, by Oddy [5]. An explicit use of the approach, however, was first introduced by Preece [6], where a mechanism called *spreading activation* was employed. Spreading activation is an algorithm for searching networks, which starts from a single node and spreads out to other nodes through edges while assigning a weight (or "activation") to each node. Although various work has been done using the spreading activation algorithm since then, it was the success of PageRank [1], which demonstrated the usefulness of the approach to the research community as well to the general public. The idea of PageRank is to utilize the hyperlink structure of Web documents, in addition to their contents, to rank retrieved documents. PageRank constructs a graph structure, where hyperlinks are represented as edges, and web documents are represented as nodes. It then assigns higher weights to nodes with more edges coming from other nodes. Thus, it effectively collects "votes" from Web pages to rank the pages.

The strength of the graph-based approach is the generalizability: the approach may be applied to any structural relations, not just the hyperlinks. Mihalcea and Tarau proposed TextRank [4], a graph-based ranking algorithm similar to PageRank, for extracting keywords from documents.

With TextRank, nodes in the graph represent words instead of Web documents, and edges represent word co-occurrences instead of hyperlinks. Coursey et al. applied the TextRank algorithm to extract keywords from learning materials of history, combined with another keyword extraction method called Wikifier [2].

4. CURRENT STUDY

4.1 Approach

Our approach is to enhance TextRank by using semantic relations, instead of term co-occurrences, as edges of the graph. Specifically, we plan to apply the definition of "semantic relatedness" using Wikipedia by Strube and Ponzetto [7]. While previous definitions of semantic relatedness have been based on word relations derived from sources such as WordNet [3], more recently researchers have used the resources available through Wikipedia, as in [7]. As a collaboratively generated corpus, Wikipedia provides a breadth and depth of topics not easily achieved otherwise. In order to define semantic relatedness, Wikipedia pages can be viewed as a collection of categorized concepts, forming a semantic network. Relations between concepts are given by a hyperlink structure between articles, forming a wide variety of relations, not just "is-a" or "part-of" relationships.

4.2 Evaluation Environment

A platform for evaluating and developing keyword extraction mechanisms has been developed using Java and CNLP's libraries that utilize the text processing engine, TextTagger, and other utilities such as html or pdf document preprocessors. The evaluation environment is depicted in Figure 1. For each learning material, the text is processed with TextTagger and the accompanying preprocessing to obtain MAST metadata. The different keyword extraction algorithms, shown as 1 through N, may use the the extracted metadata, the text directly from the learning materials and statistics from a background corpus in their definitions. These algorithms will include a baseline standard keyword algorithm, Mihalcea's TextRank algorithm, and our new algorithm based on TextRank with semantic relations. The evaluation module will compare the different keyword extraction algorithms by calculateing standard measures (precision, recall, and f-measures) based on the gold standards, which are extracted from manually created MAST metadata.

4.3 Data

The study utilizes metadata that have been created in the MAST project (described in the Background section), as the gold standard. Human annotators (information professionals) created a list of keywords that describe the document as part of the metadata. So far metadata were created for 50 online documents, which described lesson plans in mathematics and science.

5. FUTURE WORK

At the time of writing, we are implementing two keyword extraction algorithms: TextRank with word co-occurrences and TextRank with semantic relatedness. By the time of the conference, we aim to test the two algorithms against the data and present the results in the poster.

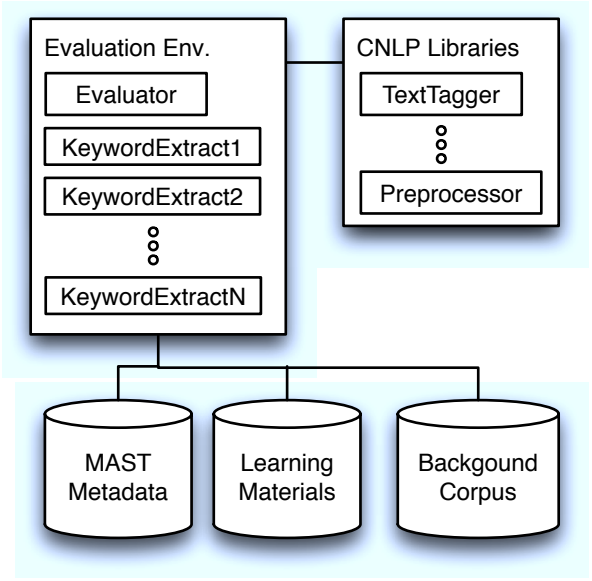

Figure 1: Evaluation Environment

6. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7), 1998.

[2] K. Coursey, R. Mihalcea, and W. Moen. Automatic keyword extraction for learning object repositories. In *Proceedings of the Conference of the American Society for Information Science and Technology*, Columbus, Ohio, October 2008.

[3] C. Fellbaum, editor. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA, 1998.

[4] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In *in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, Barcelona, Spain, July 2004.

[5] R. N. Oddy. Information retrieval through man-machine dialogue. *Journal of Documentation*, 33(1):1–14, 1977.

[6] S. E. Preece. *A spreading activation network model for information retrieval*. PhD thesis, University of Illinois at Urbana-Champaign, Champaign, IL, USA, 1981.

[7] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *The Proceedings of the Twenty-First National Conference on Artificial Intelligence*. American Association for Artificial Intelligence, July 2006.